1997 BROADCAST NEWS BENCHMARK TEST RESULTS: ENGLISH AND NON-ENGLISH

David S. Pallett, Jonathan G. Fiscus, Alvin Martin, and Mark A. Przybocki

National Institute of Standards and Technology (NIST)
Information Technology Laboratory (ITL)
Room A216 Building 225 (Technology)
Gaithersburg, MD 20899
E-mail: dpallett@nist.gov

ABSTRACT

This paper documents use of Broadcast News test materials in DARPA-sponsored Automatic Speech Recognition (ASR) Benchmark Tests conducted late in 1997. This year's English-language tests differed from last year's in that statistical selection procedures were used in selecting a three-hour test set comprised of 158 story-length segments, in contrast to last year's two-hour test set which was comprised of 4 half-hour segments. The increased number of segments is intended to provide a better statistical sampling of story-length segments and a statistically-equivalent reserve test set for a future evaluation.

The lowest word-error rate reported this year was 16.2%, contrasting with last year's lowest word error rate of 27.1%. In part, this apparent improvement is due to the much greater proportion of well-recognized F0 data present in the test set. This, in turn, is due to an effort to "balance" the test pool to match the properties of the training data.

New this year was the completion of tests in languages other than English—Mandarin and Spanish.

1. TEST MATERIALS

1.1 English Language Materials

A companion paper [1] describes the procedures used by NIST in selecting the test materials used for this year's Hub 4 English tests. This year's test set introduces statistical selection considerations, including adjustment of the properties of a test data pool so as to more accurately reflect those of the training data pool, definition of the "unit" of interest for statistical analysis as the "story," and concurrent selection of a statistically-equivalent reserve test set for a future evaluation.

Test materials were drawn from a pool of data provided by the Linguistic Data Consortium, comprising ten hours—recordings of 5 television broadcasts from 4 sources, and recordings of 4 radio broadcasts from 3 sources. These materials were supplemented by a seven hour set of recordings obtained from C-SPAN, which was used to provide "speeches"—in this case mostly from candidates for political office. Because interest had been expressed in sampling a diverse range of speeches, the sample selection algorithm, in this case, was limited to selection of fourteen one-minute excerpts, one per speaker, from the ten hours of materials.

1.2 Non-English Language Materials

For the Non-English test materials, the test set selection procedure was more constrained in that a total of five hours of potential test materials was provided by the Linguistic Data Consortium, and a one-hour test set was required. Nonetheless, the process that was followed in test data selection was similar to that used for the English materials—involving random selection of stories, and (in this case) selecting a test set that maximized the number of new speakers.

2. EVALUATION RESULTS

2.1 Evaluation Design Changes

The design of the 1997 evaluation differed in a number of ways from that of the 1996 evaluation:

- The 1997 evaluation was defined to be a "UE" evaluation (unpartitioned), whereas the 1996 evaluation included a "PE" evaluation (partitioned) component making use of time marks obtained from the hand segmentation. In 1997, NIST provided segment boundary data

using an automatic segmentation software module provided by CMU. Use of this information was optional.

- The 1997 evaluation required participating sites to process a 3-hour file consisting of a concatenation of 158 variable length excerpts spliced together, as opposed to the 1996 evaluation, which required sites to process four 1/2-hour files, each of which was chosen from a single source.

- The 1997 test data was selected so as to be representative of the training pool, whereas the 1996 test data was selected to maximize focus condition coverage.

- An additional 50 hours of acoustic training data was provided in 1997 to complement the 55 hours that was made available for the 1996 tests.

2.2 Scoring Changes

In 1996, differences existed between the scoring protocols used by NIST for Hubs 4 and 5. In 1997, a unified scoring protocol was developed and implemented. This principally involved the definition of several categories of speech artifacts (e.g. unintelligible or foreign words) as "optionally deletable." The effect of these changes was measured by rescoring the 1996 test data, using the revised scoring rules. NIST rescored four sets of results, and observed an incremental reduction in word error rate ranging from 0.8% to 1.3%.

3. PARTICIPANTS

For the English benchmark tests, this year, a total of ten research groups, from nine sites, submitted results. The groups included: Carnegie Mellon University (CMU), Cambridge University's Engineering Department ("CU-CONN" and "CU-HTK"), Dragon Systems (Dragon), GTE Internetworking's BBN Technologies (GTE/BBN), IBM's T.J. Watson Laboratories (IBM), the Oregon Graduate Institute (OGI), France's LIMSI group (LIMSI), Philips Research Laboratories (Philips), and SRI International (SRI). The groups from OGI and Philips had not participated in previous years' Hub 4 tests.

Two sites participated in the Spanish language evaluation, CMU and GTE/BBN, and two sites participated in the Mandarin language evaluation, Dragon and IBM.

4. TEST RESULTS

4.1 English

With submission of their results, sites were required to designate whether the results were for the site's "Primary" system, or for a "Contrastive" system.

Table 1 at the end of the paper indicates the word error rates obtained from the ten Primary systems. Overall, the range of reported word error rates is from a minimum of 16.2% to a maximum of 38.8%, for the complete test set comprising of 32,834 words. For the Baseline condition (F0), a minimum error rate of 9.9%, for the subset comprising 13,197 words, was achieved by the group at CU-HTK. For the Spontaneous speech focus condition (F1), the same group achieved an error rate of 15.4%, which is also the minimum for all systems for this condition.

Figure 1 illustrates the fact that spontaneous speech is more difficult than baseline speech, for all systems.

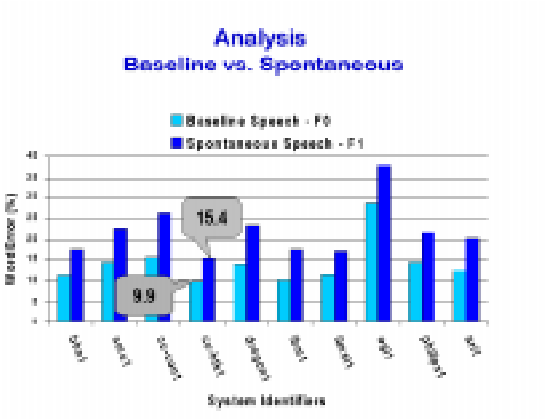

Figure 1

Table 2 at the end of the paper reports the official results of numerous two-tailed paired-comparison significance tests with the null hypothesis that there is no performance difference between the two systems.

Because this table is difficult to interpret, Figure 2 presents the results of a rank-ordered representation of overall error rates, showing the range of reported word error rates from 16.2% to 38.8%. Ovals are associated

with differences that are shown, using the Matched Pair Sentence Segment word error test [2], to have failed to reject the null hypothesis that there is no performance difference between the systems under test. Thus, the oval associated with the IBM and LIMSI results indicates that there is no significance associated with the performance difference reported for those systems (17.9% and 18.3%). Similarly, the (essentially zero) differences associated with the SRI and BBN systems are shown to be insignificant. Finally, the differences associated with the Dragon, Philips, and CMU systems (23.1%, 23.3%, and 23.8%, respectively) are shown to be of no significance.

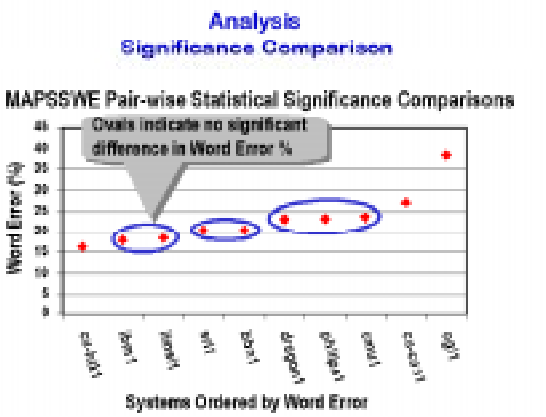

Figure 2

Implementation of the NIST-developed ROVER (Recognizer Output Voting Error Reduction) [3] system to the results reported to NIST resulted in an error rate of 12.9%.

The error rates obtained for the three individual (human) transcribers—one from the LDC, one from NIST, and one from the NSA—range from 3.3% to 4.8%.

This year, no "contrastive tests" were outlined in the test specification, but three sites submitted contrastive test results. Notable among these were results for a "near real-time" system reported by GTE/BBN, which ran in approximately 6X real-time, vs. ~200X real-time for the primary system—a 97% relative decrease in run time. For this contrastive investigation of channel and speaker normalization, a word error rate of 25.7% was measured, contrasting with 20.3% for the primary system—a 26% relative increase in word error.

A second set of contrastive test results was submitted by the Cambridge University HTK group involving the use of alternative lattice rescoring methods. These studies included replacement of a unigram cache with the NIST ROVER software, resulting in a small reduction in word error.

The third set of contrastive results submitted to NIST was from Philips and involved channel and speaker normalization as well as speaker adaptation techniques.

4.2 Non-English

There were thirty hours of training data provided for each of the two non-English languages included in this study—Spanish and Mandarin. For Spanish, the sources included Voice of America (VOA) news programs, the ECO Mexican News network, and a news program originating in Miami, Noticiero Univision. For Mandarin, sources included VOA news programs, the CCTV International China news programs (originating in Beijing), and a news program originating in Los Angeles, KAZN.

4.2.1 Spanish

As indicated previously, the test material consisted of a one hour test set selected from five hours provided by the LDC using the same selection procedure as for Hub 4 English. In this case, the test materials were drawn from the same sources as the training data. An NSA staff person was made available to NIST to verify the accuracy of the Spanish language transcripts, and to annotate the test data, so as to conform to the test specification for focus condition analysis.

For the Spanish language test materials, for CMU, the reported word error rate was 23.5%, and for GTE/BBN, it was 20.3%.

Considerable variation was observed in the degree-of-difficulty presented by the sources of test data. Word error rates for the ECO and Univision source materials ranged from between 25% to 29% for both participants, in contrast with word error rates ranging from 12% to 16% from the VOA. This variation appears to be attributable to both speaking style and rate-of-speech, since the VOA materials predominantly consist of carefully produced baseline speech.

Figure 3 shows the distribution of materials (word counts) for the three sources (ECO, Univision, and

VOA) and for the five focus conditions identified from the annotated test set (F0 - the baseline, F1 - spontaneous, F3 - speech in the presence of music, F4 - speech under degraded acoustic conditions, and FX - speech in combinations of conditions). Note that materials obtained from VOA broadcasts dominate, and of the VOA materials, the principal category is F0.

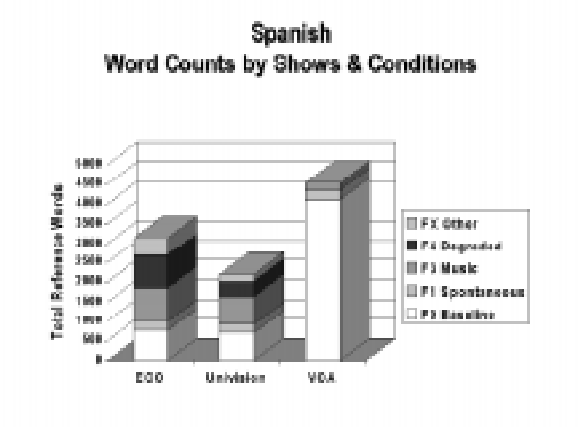

Figure 3

Figure 4 shows the corresponding word error rates obtained for the GTE/BBN system. Note that word error rates vary widely, depending on the source and condition, ranging from ~10% for baseline speech from the VOA to ~46% for spontaneous speech from Univision. These variations in the degree of difficulty point to the need for further analysis, and/or larger test sets in future non-English tests.

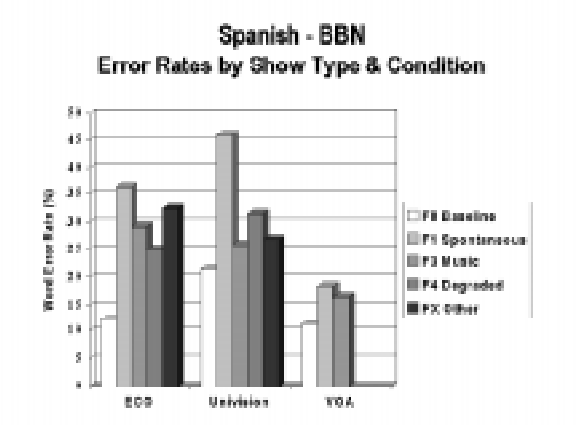

Figure 4

4.2.2 Mandarin

The test material consisted of a one hour test set developed from five hours provided by the LDC. In this case, the test materials were drawn from the same sources as the training data. Also in this case, there was no further annotation information available.

For Mandarin, scoring took place at the character level. The character error rate found the Dragon system was 20.2%, and for the IBM, 19.8%.

Figure 5 shows the Mandarin character error rate for each of the sources. Note that there is a marked difference in performance--higher error rate--for the materials originating from KAZN. These differences are probably associated with differences in the associated distribution across focus conditions—with KAZN's broadcast format consisting of AM "news radio," and having a relatively larger distribution of spontaneous speech, and of the presence of background music.

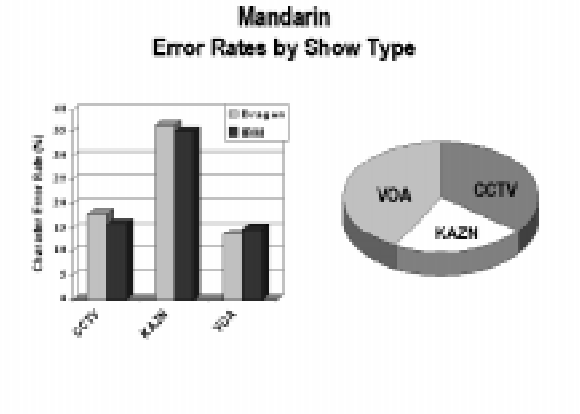

Figure 5

5. DISCUSSION

Comparing the results for those sites that participated in last year's UE tests (BBN, CMU, CU-HTK, IBM, LIMSI, and SRI) with this year's test results indicates an incremental reduction in error rate ranging from 10% to 14%. Comparing the performance of one specific system (CU-HTK) for the subsets F0 and F1 focus conditions, one finds 18.7% and 26.5% in 1996 vs. 9.9% and 15.4% in 1997.

As one of the participants noted [4], the better overall performance on this test set "seems to be due to the much greater proportion of well-recognized F0 data present." Another participant [5] noted that "the 1997 evaluation test is substantially easier than the development test set or the 1996 evaluation."

Some portion of the differences in overall performance

is undoubtedly due to the differences in the data selection paradigm used by NIST, especially our efforts to "balance" the test set with respect to the frequency-of-occurrence of materials in the different focus conditions, relying on the annotations provided by the LDC. Reconciliation of differences had the result of increasing the percentage of materials in the F0 baseline condition from 35% to 44%, and in the F1 "spontaneous" condition from 15% to 19%, so that 63% of the test set materials ended up classified in the low background noise category. However, looking at the corresponding data for 1996 [6], one finds 29.7% of that data was classified as F0, and 32.7% as F1, thus 62.4% in all in the low background noise category (almost exactly the same percentage as in 1997), so the differences that can be noted reflect greater emphasis on the F0 baseline condition—44% (in 1997) vs. 29.7% (in 1996).

6. ACKNOWLEDGMENTS

We would like to acknowledge the assistance of Audrey Le, who worked with Bill Fisher and Walter Liggett in selecting potential test materials and in analyzing the properties of the training set.

NOTICE

The views expressed in this paper are those of the authors. The test results are for local, system-developer implemented tests. NIST's role was one that involved working with the LDC in processing LDC-provided training and potential test materials, selecting and defining reference annotation and transcription files for the tests, developing and implementing scoring software, and uniformly scoring and tabulating results. The views of the authors, and these results, are not to be construed or represented as endorsements of any systems, or as official findings on the part of NIST, DARPA, or the U.S. Government.

REFERENCES

[1] Fisher, W.M., et al. "Data Selection for Broadcast News CSR Evaluations," in this Proceedings.

[2] Pallett, D.S., Fisher, W.M., and Fiscus, J.G. "Tools for the Analysis of Benchmark Speech Recognition Tests," Proceedings of ICASSP 90, pp. 97-100.

[3] Fiscus, J.G. "A Post-Processing System to Yield Reduced Word Error Rates: Recognizer Output Voting Error Reduction (ROVER)," Proceedings of the 1997 IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 347-354.

[4] Woodland, P.C., et al. "The 1997 HTK Broadcast News Transcription System," in this Proceedings.

[5] Kubala, F., et al. "The 1997 BYBLOS System Applied to Broadcast News Transcription," in this Proceedings.

[6] Garofolo, J.S., Fiscus, J.G., and Fisher, W.M. "Design and Preparation of the 1996 Hub-4 Broadcast News Benchmark Test Corpora," in Proceedings of the Speech Recognition Workshop, February 2-5, 1997, pp. 15-21.

By System Test Subset Scoring Summary
For the Hub-4E Primary Systems Test

Overall -> Overall
Baseline Broadcast Speech ->
Spontaneous Broadcast Speech ->
Speech Over Telephone Channels ->
Speech in the Presence of Background Music ->
Speech Under Degraded Acoustic Conditions ->
Speech from Non-Native Speakers ->
All other speech ->

		1996 Hub4 Focus Conditions											
SYSTEM	Overall	Baseline Broadcast Speech		Spontaneous Broadcast Speech		Speech Over Telephone Channels		Speech in the Presence of Background Music		Speech Under Degraded Acoustic Conditions		Speech from Non-Native Speakers		All other speech	
	#Wrd %WE	#Wrd	%WE	#Wrd	%WE	#Wrd	%WE	#Wrd	%WE	#Wrd	%WE	#Wrd	%WE	#Wrd	%WE
Set/Subset #Words and System Set/Subset Average Word Error Rate															
bbn1.ctm.filt	[32834] 20.3	[13197]	11.4	[6566]	17.8	[4882]	31.2	[1571]	28.1	[3350]	22.1	[669]	26.9	[2599]	42.7
cmu1.ctm.filt	[32834] 23.8	[13197]	14.4	[6566]	22.8	[4882]	31.0	[1571]	33.9	[3350]	27.3	[669]	31.1	[2599]	48.2
cu-con1.ctm.filt	[32834] 27.1	[13197]	15.5	[6566]	26.3	[4882]	37.5	[1571]	35.1	[3350]	31.2	[669]	25.7	[2599]	59.1
cu-htk1.ctm.filt	[32834] 16.2	[13197]	9.9	[6566]	15.4	[4882]	20.1	[1571]	27.9	[3350]	19.4	[669]	24.1	[2599]	29.9
dragon1.ctm.filt	[32834] 23.1	[13197]	13.9	[6566]	23.4	[4882]	31.1	[1571]	34.9	[3350]	26.5	[669]	19.0	[2599]	43.9
ibm1.ctm.filt	[32834] 17.9	[13197]	10.3	[6566]	17.8	[4882]	24.9	[1571]	24.6	[3350]	20.3	[669]	18.2	[2599]	36.3
limsi1.ctm.filt	[32834] 18.3	[13197]	11.6	[6566]	17.0	[4882]	22.1	[1571]	27.9	[3350]	21.9	[669]	27.1	[2599]	36.3
ogi1.ctm.filt	[32834] 38.8	[13197]	28.6	[6566]	38.0	[4882]	52.5	[1571]	50.0	[3350]	37.3	[669]	38.7	[2599]	62.0
philips1.ctm.filt	[32834] 23.3	[13197]	14.4	[6566]	21.7	[4882]	30.8	[1571]	34.4	[3350]	25.7	[669]	30.9	[2599]	47.1
sri1.ctm.filt	[32834] 20.3	[13197]	12.5	[6566]	20.5	[4882]	26.4	[1571]	32.0	[3350]	23.1	[669]	26.8	[2599]	35.2
Set/Subset Mean #Words/Speaker and Set/Subset Mean Word Error Rate/Speaker															
bbn1.ctm.filt	[280] 29.8	[269]	14.6	[547]	25.2	[203]	41.0	[68]	28.4	[145]	38.3	[133]	25.1	[86]	39.4
cmu1.ctm.filt	[280] 32.7	[269]	15.7	[547]	29.9	[203]	40.0	[68]	33.5	[145]	50.5	[133]	29.3	[86]	41.1
cu-con1.ctm.filt	[280] 35.2	[269]	17.2	[547]	33.1	[203]	44.0	[68]	35.3	[145]	51.2	[133]	23.3	[86]	44.9
cu-htk1.ctm.filt	[280] 23.8	[269]	10.6	[547]	23.2	[203]	26.7	[68]	30.1	[145]	35.7	[133]	23.0	[86]	32.9
dragon1.ctm.filt	[280] 31.2	[269]	14.9	[547]	30.4	[203]	41.1	[68]	33.6	[145]	37.2	[133]	20.8	[86]	42.7
ibm1.ctm.filt	[280] 25.6	[269]	11.9	[547]	22.2	[203]	32.7	[68]	27.1	[145]	33.4	[133]	16.0	[86]	35.1
limsi1.ctm.filt	[280] 24.4	[269]	11.9	[547]	23.6	[203]	28.5	[68]	31.2	[145]	34.4	[133]	24.7	[86]	35.2
ogi1.ctm.filt	[280] 47.5	[269]	32.4	[547]	49.5	[203]	57.5	[68]	46.7	[145]	56.1	[133]	37.4	[86]	53.7
philips1.ctm.filt	[280] 32.4	[269]	16.4	[547]	29.1	[203]	40.0	[68]	35.5	[145]	43.6	[133]	29.4	[86]	42.0
sri1.ctm.filt	[280] 27.3	[269]	14.7	[547]	30.4	[203]	33.4	[68]	29.6	[145]	35.4	[133]	27.6	[86]	35.4
Associated Standard Deviations															
bbn1.ctm.filt	[410] 23.9	[288]	12.9	[729]	13.5	[308]	23.5	[65]	21.1	[170]	31.0	[76]	8.3	[76]	25.3
cmu1.ctm.filt	[410] 29.3	[288]	15.0	[729]	15.3	[308]	27.1	[65]	25.7	[170]	43.5	[76]	8.3	[76]	23.8
cu-con1.ctm.filt	[410] 27.6	[288]	12.1	[729]	10.8	[308]	23.9	[65]	23.4	[170]	43.6	[76]	9.2	[76]	23.7
cu-htk1.ctm.filt	[410] 20.5	[288]	9.1	[729]	18.2	[308]	19.4	[65]	21.2	[170]	29.6	[76]	4.5	[76]	21.1
dragon1.ctm.filt	[410] 22.0	[288]	11.4	[729]	16.0	[308]	21.0	[65]	20.5	[170]	24.9	[76]	7.7	[76]	20.7
ibm1.ctm.filt	[410] 22.7	[288]	14.2	[729]	10.8	[308]	20.4	[65]	15.0	[170]	26.9	[76]	8.1	[76]	23.5
limsi1.ctm.filt	[410] 20.3	[288]	9.6	[729]	16.7	[308]	19.7	[65]	22.7	[170]	27.6	[76]	8.2	[76]	24.3
ogi1.ctm.filt	[410] 23.3	[288]	20.2	[729]	16.7	[308]	23.0	[65]	21.7	[170]	26.9	[76]	8.1	[76]	22.6
philips1.ctm.filt	[410] 27.4	[288]	12.5	[729]	14.5	[308]	27.2	[65]	17.3	[170]	40.6	[76]	10.4	[76]	24.1
sri1.ctm.filt	[410] 21.3	[288]	11.1	[729]	17.0	[308]	22.4	[65]	21.0	[170]	29.5	[76]	4.2	[76]	22.4
Set/Subset Median #Words/Speaker and Set/Subset Median Word Error Rate/Speaker															
bbn1.ctm.filt	[129] 21.2	[174]	10.2	[182]	23.5	[61]	31.3	[46]	22.5	[64]	25.0	[129]	25.2	[73]	33.4
cmu1.ctm.filt	[129] 24.3	[174]	11.6	[182]	23.7	[61]	31.2	[46]	28.6	[64]	37.8	[129]	30.5	[73]	36.6
cu-con1.ctm.filt	[129] 29.6	[174]	13.4	[182]	30.4	[61]	36.3	[46]	28.6	[64]	35.8	[129]	23.2	[73]	41.8
cu-htk1.ctm.filt	[129] 17.4	[174]	8.3	[182]	15.4	[61]	22.2	[46]	22.4	[64]	31.1	[129]	23.4	[73]	27.3
dragon1.ctm.filt	[129] 25.2	[174]	12.5	[182]	23.2	[61]	33.1	[46]	29.4	[64]	32.1	[129]	17.8	[73]	42.8
ibm1.ctm.filt	[129] 19.8	[174]	7.9	[182]	18.8	[61]	25.5	[46]	24.3	[64]	28.9	[129]	16.3	[73]	25.3
limsi1.ctm.filt	[129] 16.4	[174]	10.6	[182]	16.3	[61]	21.4	[46]	24.7	[64]	26.7	[129]	25.9	[73]	29.4
ogi1.ctm.filt	[129] 45.3	[174]	24.1	[182]	46.7	[61]	54.0	[46]	47.1	[64]	50.0	[129]	36.4	[73]	52.4
philips1.ctm.filt	[129] 26.3	[174]	12.6	[182]	22.6	[61]	33.1	[46]	35.0	[64]	33.3	[129]	27.9	[73]	36.4
sri1.ctm.filt	[129] 21.2	[174]	11.7	[182]	27.0	[61]	25.8	[46]	26.6	[64]	24.1	[129]	28.0	[73]	31.5

Table 1

Composite Report of All Significance Tests
For the Hub-4E Primary Systems Test Test

Test Name	Abbrev.
Matched Pair Sentence Segment (Word Error)	MP
Signed Paired Comparison (Speaker Word Error Rate (%))	SI
Wilcoxon Signed Rank (Speaker Word Error Rate (%))	WI
McNemar (Sentence Error)	MN

Test Abbrev.		bbn1	cmu1	cu-con1	cu-htk1	dragon1	ibm1	limsi1	ogi1	philips1	sri1	Test Abbrev.
MP	bbn1		bbn1	bbn1	cu-htk1	bbn1	ibm1	limsi1	bbn1	bbn1	~	MP
SI			bbn1	bbn1	cu-htk1	bbn1	ibm1	limsi1	bbn1	bbn1	~	SI
WI			bbn1	bbn1	cu-htk1	bbn1	ibm1	limsi1	bbn1	bbn1	~	WI
MN			bbn1	bbn1	cu-htk1	~	ibm1	limsi1	bbn1	~	sri1	MN
MP	cmu1			cmu1	cu-htk1	~	ibm1	limsi1	cmu1	~	sri1	MP
SI				cmu1	cu-htk1	~	ibm1	limsi1	cmu1	~	sri1	SI
WI				cmu1	cu-htk1	~	ibm1	limsi1	cmu1	~	sri1	WI
MN				~	cu-htk1	dragon1	ibm1	limsi1	cmu1	philips1	sri1	MN
MP	cu-con1				cu-htk1	dragon1	ibm1	limsi1	cu-con1	philips1	sri1	MP
SI					cu-htk1	dragon1	ibm1	limsi1	cu-con1	philips1	sri1	SI
WI					cu-htk1	dragon1	ibm1	limsi1	cu-con1	philips1	sri1	WI
MN					cu-htk1	dragon1	ibm1	limsi1	cu-con1	philips1	sri1	MN
MP	cu-htk1					cu-htk1	cu-htk1	cu-htk1	cu-htk1	cu-htk1	cu-htk1	MP
SI						cu-htk1	~	cu-htk1	cu-htk1	cu-htk1	cu-htk1	SI
WI						cu-htk1	cu-htk1	cu-htk1	cu-htk1	cu-htk1	cu-htk1	WI
MN						cu-htk1	~	~	cu-htk1	cu-htk1	~	MN
MP	dragon1						ibm1	limsi1	dragon1	~	sri1	MP
SI							ibm1	limsi1	dragon1	~	sri1	SI
WI							ibm1	limsi1	dragon1	~	sri1	WI
MN							ibm1	limsi1	dragon1	~	sri1	MN
MP	ibm1							~	ibm1	ibm1	ibm1	MP
SI								~	ibm1	ibm1	ibm1	SI
WI								~	ibm1	ibm1	ibm1	WI
MN								~	ibm1	ibm1	~	MN
MP	limsi1								limsi1	limsi1	limsi1	MP
SI									limsi1	limsi1	limsi1	SI
WI									limsi1	limsi1	limsi1	WI
MN									limsi1	limsi1	~	MN
MP	ogi1									philips1	sri1	MP
SI										philips1	sri1	SI
WI										philips1	sri1	WI
MN										philips1	sri1	MN
MP	philips1										sri1	MP
SI											sri1	SI
WI											sri1	WI
MN											sri1	MN
MP	sri1											MP
SI												SI
WI												WI
MN												MN

These significance tests are all two-tailed tests with the null hypothesis
that there is no performance difference between the two systems.

Table 2